A Child's Play: An Agent-based Simulator to Protect Minors Online

Keywords: Agents, Artificial Intelligence, Human behavior modeling, Interactive story-like serious games, Synthetic data generation

Extended Abstract

One in three Internet users were minors in 2017, based on UNICEF report "Children in a Digital World" [1]. The COVID-19 pandemic has enlarged such a phenomenon, making more minors spend more time online at younger ages. Consequently, they are also more exposed to online threats. In an interdisciplinary effort bringing together LEA (Law Enforcement Agencies), psychologists, sociologists, educators, lawyers, experts in ethics, and computer scientists, our team set out to fight cybercrime by developing an evidence-based interactive story-like serious game [2] (henceforth, *game*). With this *game*, we aim to teach good practices and identify potential risk indicators based on the data gathered through it, so that LEA can develop science-based policies to improve minors' online experience.

Our pursuit faces two main challenges, which are usually faced by other social experiments: the scarcity of the data (or variability within it) and the limitations of the instruments to discriminate among causal and spurious correlations.

In this context, Bayesian generative methods help to address both difficulties (Figure 1, top panel). On the one hand, they incorporate prior knowledge on the field (which also helps assess its validity), creating a (probability-based) generative model in which the expert's knowledge can be translated into a hierarchy of relevant variables, capturing the causal relationship among variables. The side benefit of this approach is that scarce or imbalanced data (for instance, low prevalence of some social or psychological profiles) can still be analyzed quantitatively in the Bayesian framework.

To deal with the data scarcity in particular, artificial intelligence can be combined with Bayesian generative methods to generate synthetic data. However, contrary to traditional approaches, which train machines with the objective of winning the game [3], our goal is to mimic the behavior of our participants (minors) while playing such a *game* related to cybercrime and online risk, in an approach more similar to [4,5]. Such synthetic data can, in turn, be helpful for game data analytics, compensating the potentially harmful effects of imbalanced datasets, allowing training, evaluating, selecting, and calibrating ML (Machine Learning) algorithms, or enabling data augmentation.

Figure 1 (bottom panel) shows an overview of the framework proposed to simulate human behavior and generate synthetic data. Such a framework receives the *game* script in a given format and obtains the graph representation of the story, translating it also to HTML (HypeText Markup Language). An API (Application Programming Interface) is defined so that different agents simulating different behaviors can interact with the HTML version of the *game* uniformly. The game storyline is about a teenager's life during five days of his/her life in which he/she is exposed to online grooming. The game puts the player in different risky situations where past (poor) decisions affect future outcomes.

The target generative model has to be flexible enough to accommodate any expected behavior of the players. Thus, we will implement a wide variety of *agents*: random player, conservative player, risky player, short-lived risky player, explorer/blackmailed player.

The random player is not the only one who acts *stochastically*. Each of the other players has a *propensity* to behave as expected by its role, while also choosing its actions stochastically. Moreover, different machine learning approaches can be used to model these behaviors, such as agents based on neural networks, reinforcement learning, finite state machines, heuristics/profiles (i.e., rule-based), or behavioral trees (i.e., hierarchical rule-based).

To automate choice labeling based on text, we tested two architectures (motivated by the idiosyncrasies of natural language): LSTM (Long Short-Term Memory) [6] and Fine-tuned BERT (Bidirectional Encoder Representations from Transformers) classifiers [7].

In the preliminary tests we have carried out with the aforementioned *game*, the LSTM model obtains an accuracy of 0.54 on the validation set and 0.63 on the test set. The fine-tuned BERT obtains an accuracy of 0.75 on the validation set and 0.71 on the test set. The upper bound in the accuracy (around 0.75) is caused both by the lack of representative risky choices and the ambiguity of natural language (see Figure 2). Although there is still some room for improvement, an accuracy score of 0.75 is considerable in the Social Sciences; adult humans might get closer to 100%, but children and adolescents might not be far from that number.

Inspection of the table in Figure 2C shows how our model has learned from the *game* that talking about the experience to “your loved ones,” such as your mother or a friend, is a good decision and, not doing so, is a potentially dangerous decision. Moreover, the architecture measures the uncertainty of the outcome: family-related topics are more informative than “stranger”-based content. We believe the main reason this happens is that the situations presented in the *game* are not complex enough to learn this distinction correctly. One solution for this issue might be using a *game* with more complex interactions and a greater variety of decisions.

In summary, we present a prototype framework that combines a custom-made interactive story-like serious game, non-parametric Bayesian modeling (where the generative part of the model is an Agent-based simulation) with a Natural Language Processing (NLP) Deep Learning architecture to label *risky* decisions. All in all, we have built an inferential model to categorize real (minor) players upon interaction with an *appealing* game and, more promisingly, to detect risky patterns that can inform educational interventions.

References

[1] Children in a Digital World. UNICEF (2017). https://www.unicef.org/media/48601/file

[2] Abt, C. C. Serious Games. New York: Viking (1970).

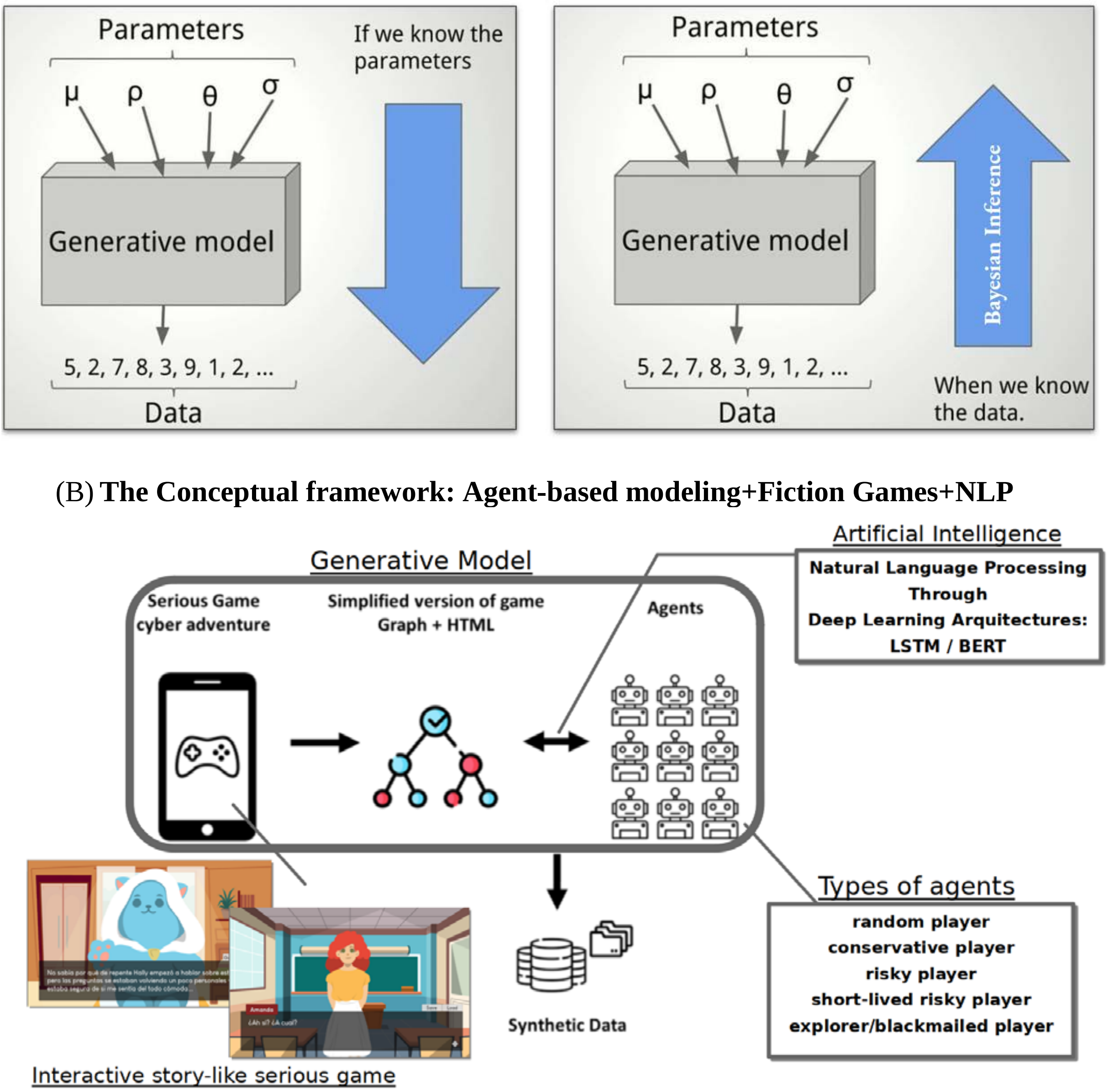
[3] Silver, D., Huang, A., Maddison, C. *et al.* Mastering the game of Go with deep neural networks and tree search. *Nature* 529, 484–489 (2016). https://doi.org/10.1038/nature16961.

[4] Wang, B., Sun, T., Zheng, X. S. Beyond Winning and Losing: Modeling Human Motivations and Behaviors with Vector-Valued Inverse Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence and Interactive Digital Entertainment, 15*(1), 195-201 (2019). https://ojs.aaai.org/index.php/AIIDE/article/view/5244.

[5] Lin, B., Cecchi, G., Bouneouf, D., Reinen, J., Rish, I. A Story of Two Streams: Reinforcement Learning Models from Human Behavior and Neuropsychiatry. *Proceedings of the Nineteenth International Conference on Autonomous Agents and Multi-Agent Systems,* 744-752 (2020). https://ifaamas.org/Proceedings/aamas2020/pdfs/p744.pdf

[6] Hochreiter, S., Schmidhuber, J. Long short-term memory. *Neural Computation* 9(8) 1735-1780 (1997). https://doi.org/10.1162/neco.1997.9.8.1735

[7] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1,* 4171–4186 (2019). https://aclanthology.org/N19-1423/

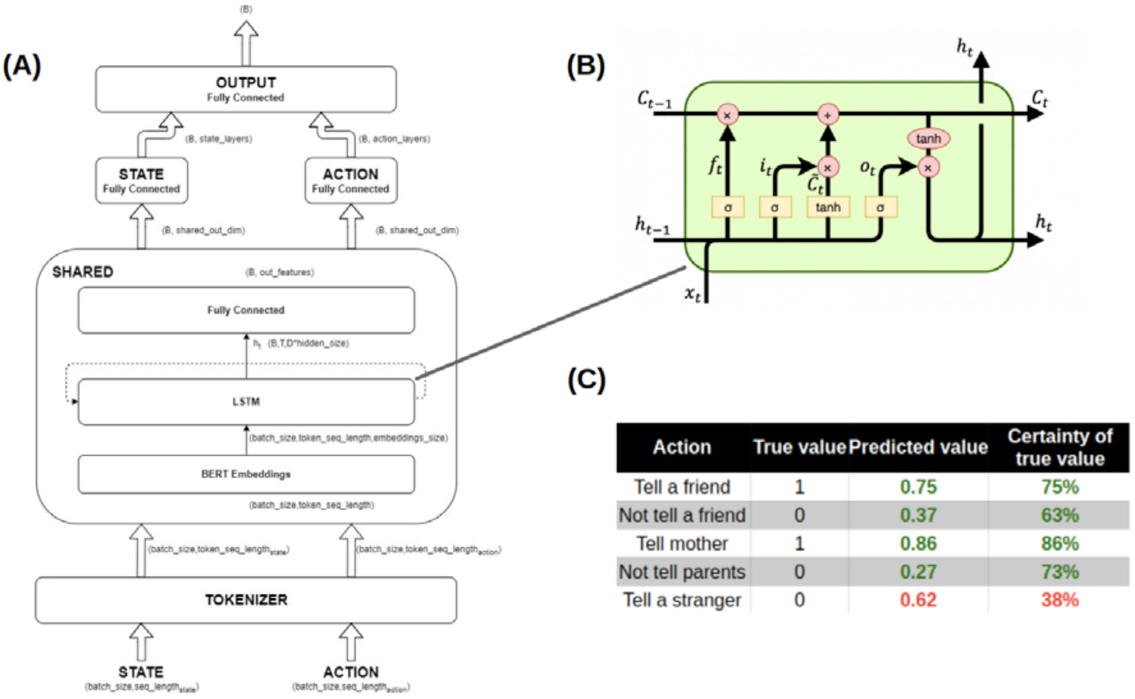


(A) The Bayesian generative approach

(B) The Conceptual framework: Agent-based modeling+Fiction Games+NLP

Generative Model
Serious Game cyber adventure
Simplified version of game Graph + HTML
Agents
Artificial Intelligence
Natural Language Processing Through Deep Learning Arquitectures: LSTM / BERT
Synthetic Data
Types of agents
random player
conservative player
risky player
short-lived risky player
explorer/blackmailed player
Interactive story-like serious game

Figure 1. (A) Non-parametric Bayesian models can be viewed as a black-box that *generates* data according to some pre-establish algorithm (in our case, Agent-Based model). Upon contrasting the synthetic data with real-life observations, the model can estimate latent variables such as “propensity to risk” or “credibility of the responses” of the players in the serious game. (B) Overview of the proposed framework to simulate human behavior and generate synthetic data. Agents can fall into any of the following categories: (i) random player (as we expect some minors to be bored during the task and type randomly); (ii) safe or conservative player (the agent identifies safer alternatives and choose them consistently); (iii) risky player (the opposite to the safe player); (iv), short-lived risky player (e.g., a player who starts being very risky but evolves to more conservative decisions while playing, thus involving learning from previous bad experiences), (v) explorer/blackmailed player (e.g., if it chooses poorly once, this may lead to taking more and more risky decisions).



Action	True value	Predicted value	Certainty of true value
Tell a friend	1	0.75	75%
Not tell a friend	0	0.37	63%
Tell mother	1	0.86	86%
Not tell parents	0	0.27	73%
Tell a stranger	0	0.62	38%

Figure 2. (A) Natural-Language Processing module of the architecture in Figure 1B. The model takes the scene in the *game* to parse the relevant information and determine the risk of taking different actions from the current state (B) Example of an LSTM neuron showing the recurrent nature of the architecture. Our model was programmed in the Python programming language using the PyTorch library. (C) Table containing 5 randomly chosen examples from the test set. Given that the game does not carry sufficient information to train a model from scratch in many cases, both classifiers use a pre-trained BERT model. The model provides two pieces of information: the *predicted variable* and, interestingly, the *certainty* about the reported value. We can observe how our model has learned from the *game* that talking about the experience to your loved ones, such as your mother or a friend, is a good decision and not doing so is a wrong decision. When the action is referred to the family, it is even more certain about its decision. However, when presented with a similar situation and an adverse action (*tell a stranger*), we can observe how the model is not as confident as before but still decides to tell the stranger. The model has learned accurately that it is right to be open about your problems and talk about them, but it has not learned to filter to whom it is right to do so correctly.